Interspeech 2018
2-6 September 2018, Hyderabad

Role Play Dialogue Aware Language Models based on Conditional Hierarchical Recurrent Encoder-Decoder

Ryo Masumura, Tomohiro Tanaka, Atsushi Ando, Hirokazu Masataki, Yushi Aono

NTT Media Intelligence Laboratories, NTT Corporation, Japan

ryou.masumura.ba@hco.ntt.co.jp

Abstract

We propose role play dialogue-aware language models (RPDA-LMs) that can leverage interactive contexts in role play multi-turn dialogues for estimating the generative probability of words. Our motivation is to improve automatic speech recognition (ASR) performance in role play dialogues such as contact center dialogues and service center dialogues. Although long short-term memory recurrent neural network based language models (LSTM-RNN-LMs) can capture long-range contexts within an utterance, they cannot utilize sequential interactive information between speakers in multi-turn dialogues. Our idea is to explicitly leverage speakers' roles of individual utterances, which are often available in role play dialogues, for neural language modeling. The RPDA-LMs are represented as a generative model conditioned by a role sequence of a target role play dialogue. We compose the RPDA-LMs by extending hierarchical recurrent encoder-decoder modeling so as to handle the role information. Our ASR evaluation in a contact center dialogue demonstrates that RPDA-LMs outperform LSTM-RNN-LMs and document-context LMs in terms of perplexity and word error rate. In addition, we verify the effectiveness of explicitly taking interactive contexts into consideration.

Index Terms: role play dialogue aware language models, hierarchical recurrent encoder-decoder, automatic speech recognition, contact center dialogues

1. Introduction

In the automatic speech recognition (ASR) field, multi-party conversation is one of the most popular ASR tasks. Multi-party conversation tasks have different attributes from typical single speaker tasks since multiple speakers interact with each other. In order to enhance ASR performance in multi-party conversation tasks, language models (LMs) have to precisely capture the interactive contexts.

A lot of studies have been reported for improving language modeling in single speaker tasks. For a while, smoothed n-gram LMs were employed in ASR because they yield powerful performance in spite of simple modeling [1–4]. In recent studies, neural LMs that capture words by converting continuous representations have attracted a lot of attention [5, 6]. In particular, long short-term memory recurrent neural network based LMs (LSTM-RNN-LMs) provide effective modeling to leverage long-range sequential contexts within an utterance [7–9]. In addition, to capture discourse-level sequential contexts, document context LMs that can capture long-range sequential contexts beyond utterance boundaries have also been proposed [10, 11].

On the other hand, language models intended to be used in multi-party conversation have also been developed [12, 13]. In order to capture differences between speakers, speakers' role information is often utilized for language modeling [14, 15]. In fact, the role information is easily obtained in role play dialogue ASR tasks such as contact center dialogues or service center dialogues because individual speaker's speech can be recorded on pre-defined different sources. Most previous work utilized the role information for reflecting role-specific word occurrences into n-gram LMs [15].

However, the previous work did not take sequential interactive contexts between speakers into consideration. We can expect that ASR performance can be improved by capturing the sequential interactive contexts since they incrementally affect the next speaker's utterances. Therefore, our idea is to reconcile role-dependent modeling with LSTM-RNN based long-range sequential modeling. Although a similar idea is examined in dialogue context LMs, they can only be applied for two-speaker dialogues in which each speaker's utterances are uttered alternately [16]. In real ASR applications, more flexible LMs are required because two speaker's utterances rarely alternate.

In this paper, we propose role play dialogue aware LMs (RPDA-LMs) that can leverage sequential interactive contexts from start-of-conversation to a current word for estimating the generative probability of a word in a current utterance. Unlike the conventional dialogue context LMs, the RPDA-LMs can be applied to arbitrary role play dialogue ASR tasks. We compose the RPDA-LMs by extending a hierarchical recurrent encoder-decoder so as to handle role information [17]. In addition to the RPDA-LMs, this paper presents role-aware LSTM-RNN-LMs that can only handle role information in the current utterance in order to clarify the differences between the RPDA-LMs and the LSTM-RNN-LMs.

The RPDA-LMs are closely related to neural conversation models that are developed for a response generation module in spoken dialogue systems [18–22]. The neural conversation models are regarded as LMs conditioned on collocutor's utterances. Furthermore, dialogue context neural conversation models have been examined for incorporating multi-turn dialogues [23–25]. The RPDA-LMs are regarded as the dialogue context neural conversational models that are modified so as to accommodate role play dialogue ASR tasks. To the best of our knowledge, this paper is the first work on applying the neural conversation models to ASR. Additionally, the RPDA-LMs are related to neural language models conditioned by auxiliary features except for words. Topic information is usually introduced for capturing global lexical context [26–28]. In addition, speech information is used as the auxiliary features for incorporating a speaker's attributes [29]. In the RPDA-LMs and the role-aware LSTM-RNN-LMs, the auxiliary features are role information. Although neural networks that depend on role information have been examined in spoken language understanding fields [30, 31], this paper reports the initial study of role dependent neural networks in language modeling.

Our evaluation uses Japanese contact center dialogue data sets. We demonstrate the RPDA-LMs outperform the role-

10.21437/Interspeech.2018-2185

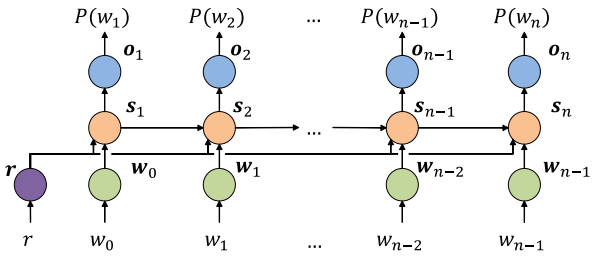

Figure 1: *Model structure of RA-LSTM-RNN-LMs.*

aware LSTM-RNN-LMs and document-context LMs in terms of perplexity and word error rate. In addition, we verify the effectiveness of explicitly taking sequential interactive contexts into consideration.

2. Role Aware LSTM-RNN LMs

This section describes role aware LSTM-RNN LMs (RA-LSTM-RNN-LMs) that can leverage role information for estimating a generative probability of words. The RA-LSTM-RNN-LMs are composed by extending LSTM-RM-LMs that can flexibly take into consideration long-range context based on an LSTM-RNN. While the LSTM-RNN-LMs are expressed as generative models, The RA-LSTM-RNN-LMs are expressed as conditional generative models. In the RA-LSTM-RNN-LMs, a role label of a target utterance is treated as given information. This modeling assumes that all utterances are mutually independent from each other. A generative probability of words $W = \{w_1, \cdots, w_N\}$ given a role label r is formulated as:

$$\begin{aligned} P(W|r, \boldsymbol{\Theta}) &= \prod_{n=1}^{N} P(w_n|w_1, \cdots, w_{n-1}, r, \boldsymbol{\Theta}) \\ &= \prod_{n=1}^{N} P(w_n|\boldsymbol{s}_n, \boldsymbol{\Theta}), \end{aligned} \tag{1}$$

where $\boldsymbol{\Theta}$ represents the model parameter and $\boldsymbol{s}_n$ is a continuous representation that embeds both $\{w_1, \cdots, w_{n-1}\}$ and r on the basis of neural network based modeling.

Figure 1 shows the model structure of the RA-LSTM-RNN-LMs. In the RA-LSTM-RNN-LMs, individual words and a role label in an utterance are converted into a continuous representation. The continuous representations of w_{n-1} and r are defined as:

$$\boldsymbol{w}_{n-1} = \texttt{EMBED}(w_{n-1}; \boldsymbol{\theta}_{\texttt{w}}), \tag{2}$$

$$\boldsymbol{r} = \texttt{EMBED}(r; \boldsymbol{\theta}_{\texttt{r}}), \tag{3}$$

where $\texttt{EMBED}()$ is a linear transformational function to embed a symbol into a continuous vector and $\boldsymbol{\theta}_{\texttt{w}}$ and $\boldsymbol{\theta}_{\texttt{r}}$ are trainable parameters.

While the LSTM-RNN-LMs capture only the word continuous representation, the RA-LSTM-RNN-LMs handle both the word continuous representation and the role continuous representation. To this end, $\boldsymbol{w}_n$ and $\boldsymbol{r}$ are merged as:

$$\boldsymbol{c}_{n-1} = [\boldsymbol{w}_{n-1}{}^{\top}, \boldsymbol{r}^{\top}]^{\top}. \tag{4}$$

The merged representation is converted into a hidden representation that summarizes past context information using an LSTM-RNN. The hidden representation for estimating the n-th word is calculated as:

$$\begin{aligned} \boldsymbol{s}_n &= \texttt{LSTM}(\boldsymbol{c}_1, \cdots, \boldsymbol{c}_{n-1}, \boldsymbol{\theta}_{\texttt{s}}) \\ &= \texttt{LSTM}(\boldsymbol{c}_{n-1}, \boldsymbol{s}_{n-1}, \boldsymbol{\theta}_{\texttt{s}}), \end{aligned} \tag{5}$$

where $\texttt{LSTM}()$ is a function of the unidirectional LSTM-RNN layer, $\boldsymbol{\theta}_{\texttt{h}}$ is the trainable parameter, $\boldsymbol{s}_0$ is a zero vector $\mathbf{0}$, and w_0 means an initial symbol. In an output layer, predicted probabilities are produced by:

$$\boldsymbol{o}_n = \texttt{SOFTMAX}(\boldsymbol{s}_n, \boldsymbol{\theta}_{\texttt{o}}), \tag{6}$$

where $\texttt{SOFTMAX}()$ is a transformational function with softmax activation, $\boldsymbol{\theta}_{\texttt{o}}$ is the trainable parameter, and $\boldsymbol{o}_n$ means $P(w_n|w_1, \cdots, w_{n-1}, r, \boldsymbol{\Theta})$. In this modeling, $\boldsymbol{\Theta}$ corresponds to $\{\boldsymbol{\theta}_{\texttt{w}}, \boldsymbol{\theta}_{\texttt{r}}, \boldsymbol{\theta}_{\texttt{s}}, \boldsymbol{\theta}_{\texttt{o}}\}$. The parameter can be optimized by:

$$\hat{\boldsymbol{\Theta}} = \operatorname*{argmin}_{\boldsymbol{\Theta}} - \sum_{(W,r)\in\mathcal{D}} \log P(W|r, \boldsymbol{\Theta}), \tag{7}$$

where $\mathcal{D}$ denotes the training data set. The optimization is conducted using backpropagation through time.

3. Role-Play Dialogue Aware LMs

This section describes role play dialogue aware LMs (RPDA-LMs). The RPDA-LMs can take relationships between utterances into consideration while the RA-LSTM-RNN-LMs can consider only words within an utterance. The RPDA-LMs compute the generative probability of words in the target utterances using not only role labels in them but also words and role labels in past utterances.

The RPDA-LMs are modeled by extending hierarchical recurrent encoder-decoder modeling so as to handle role information. In this modeling, context information in past utterances is encoded into a continuous representation in two stages. In the first stage, utterance-level information, which includes a role label and words, is converted into a continuous representation. In the second stage, continuous representations of all past utterances are also converted into a continuous representation. The generative probability of a word in a target utterance is estimated by a continuous representation that embeds the continuous representation of past utterances, a role label of a target utterance and past words in the target utterance. The words in a multi-turn dialogue is defined as $\boldsymbol{W} = \{W^1, \cdots, W^T\}$ where $W^t = \{w_1^t, \cdots, w_{N^t}^t\}$ represents the t-th utterance and N^t represents the number of words in the t-th utterance. The generative probability of $\boldsymbol{W}$ is defined as:

$$\begin{aligned} &P(\boldsymbol{W}|\boldsymbol{R}, \boldsymbol{\Theta}) \\ &= \prod_{t=1}^{T} P(W^t|W^1, \cdots, W^{t-1}, r^1, \cdots, r^t, \boldsymbol{\Theta}) \\ &= \prod_{t=1}^{T} \prod_{n=1}^{N_t} P(w_n^t|w_1^t, \cdots, w_{n-1}^t, r^t, \\ &\qquad W^1, \cdots, W^{t-1}, r^1, \cdots, r^{t-1}, \boldsymbol{\Theta}) \\ &= \prod_{t=1}^{T} \prod_{n=1}^{N_t} P(w_n^t|w_1^t, \cdots, w_{n-1}^t, r^t, \boldsymbol{u}^1, \cdots, \boldsymbol{u}^{t-1}, \boldsymbol{\Theta}) \\ &= \prod_{t=1}^{T} \prod_{n=1}^{N_t} P(w_n^t|w_1^t, \cdots, w_{n-1}^t, r^t, \boldsymbol{h}^t, \boldsymbol{\Theta}) \\ &= \prod_{t=1}^{T} \prod_{n=1}^{N_t} P(w_n^t|\boldsymbol{s}_n^t, \boldsymbol{\Theta}), \end{aligned} \tag{8}$$

where $\boldsymbol{R} = \{r^1, \cdots, r^T\}$ is a role sequence in the multi-turn dialogue which corresponds to $\boldsymbol{W}$, $\boldsymbol{\Theta}$ represents the

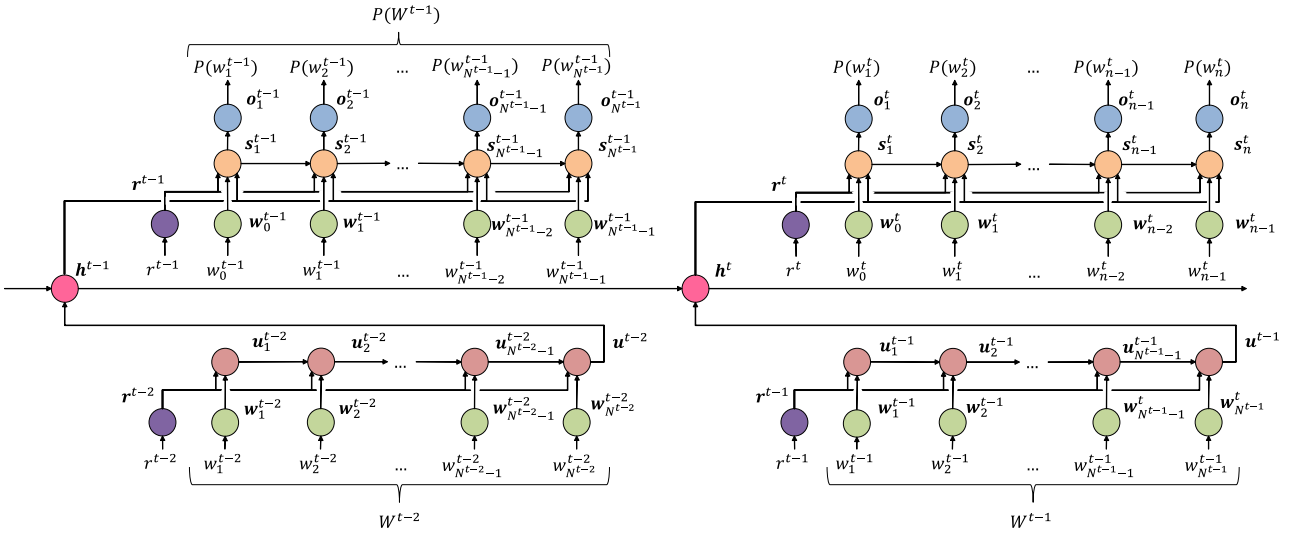

Figure 2: *Model structure of RPDA-LMs.*

model parameter, $\boldsymbol{u}^{t-1}$ is a continuous representation that embeds $\{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_{n-1}\}$ and $\{r_1, \cdots, r_{n-1}\}$, $\boldsymbol{h}^t$ is a continuous representation that embeds $\{W^1, \cdots, W^{t-1}\}$ and $\{r^1, \cdots, r^{t-1}\}$, and $\boldsymbol{s}^t_{n-1}$ is a continuous representation that embeds $\{w^t_1, \cdots, w^t_{n-1}\}$, r^t and all previous interactive contexts.

Figure 2 shows the model structure of the RPDA-LMs. The RPDA-LMs are constructed from a word-level encoder network, an utterance-level encoder network, and a decoder network. In the RPDA-LMs, each word and each role is converted into continuous representations followed by:

$$\boldsymbol{w}^t_n = \texttt{EMBED}(w^t_n; \boldsymbol{\theta}_{\texttt{w}}), \tag{9}$$

$$\boldsymbol{r}^t = \texttt{EMBED}(r^t; \boldsymbol{\theta}_{\texttt{r}}), \tag{10}$$

These continuous representations are used in both the word-level encoder network and the decoder network.

In the word-level encoder network, all words in an utterance is embedded into a continuous representation followed by:

$$\boldsymbol{z}^{t-1}_n = [\boldsymbol{w}^{t-1\top}_n, \boldsymbol{r}^{t-1\top}]^\top, \tag{11}$$

$$\begin{aligned}\boldsymbol{u}^{t-1}_n &= \texttt{LSTM}(\boldsymbol{z}^{t-1}_1, \cdots, \boldsymbol{z}^{t-1}_n, \boldsymbol{\theta}_{\texttt{u}}) \\ &= \texttt{LSTM}(\boldsymbol{z}^{t-1}_n, \boldsymbol{u}^{t-1}_{n-1}, \boldsymbol{\theta}_{\texttt{u}}),\end{aligned} \tag{12}$$

where $\boldsymbol{\theta}_{\texttt{u}}$ is the trainable parameter. The entire utterance information can be embedded into $\boldsymbol{u}^{t-1}_{N^{t-1}}$, which is expressed as:

$$\boldsymbol{u}^{t-1} = \boldsymbol{u}^{t-1}_{N^{t-1}}. \tag{13}$$

In addition, in order to capture multi-turn dialogue context information, continuous representations of past utterances are embedded into a continuous representation using the utterance-level decoder network. A continuous representation that embeds a start-of-dialogue and the $t-1$-th utterance is defined as:

$$\begin{aligned}\boldsymbol{h}^t &= \texttt{LSTM}(\boldsymbol{u}^1, \cdots, \boldsymbol{u}^{t-1}, \boldsymbol{\theta}_{\texttt{h}}) \\ &= \texttt{LSTM}(\boldsymbol{u}^{t-1}, \boldsymbol{h}^{t-1}, \boldsymbol{\theta}_{\texttt{h}}),\end{aligned} \tag{14}$$

where $\boldsymbol{\theta}_{\texttt{h}}$ is the trainable parameter. Note that $\boldsymbol{u}^0$ and $\boldsymbol{h}^0$ represent a zero vector $\mathbf{0}$. If $\boldsymbol{h}^{t-1}$ is set as a zero vector $\mathbf{0}$, only the $t-1$-th utterance is embeded into $\boldsymbol{h}^t$.

In the decoder network, which corresponds to a conditional generative model, a word continuous representation, a role continuous representation, and a continuous representation of past utterances are merged as:

$$\boldsymbol{c}^t_{n-1} = [\boldsymbol{h}^{t\top}, \boldsymbol{w}^{t\top}_{n-1}, \boldsymbol{r}^{t\top}]^\top. \tag{15}$$

The merged representation is converted into a hidden representation that summarizes all past context using an LSTM-RNN. The hidden representation for estimating the n-th word is calculated as:

$$\begin{aligned}\boldsymbol{s}^t_n &= \texttt{LSTM}(\boldsymbol{c}^t_1, \cdots, \boldsymbol{c}^t_{n-1}, \boldsymbol{\theta}_{\texttt{s}}) \\ &= \texttt{LSTM}(\boldsymbol{c}^t_{n-1}, \boldsymbol{s}^t_{n-1}, \boldsymbol{\theta}_{\texttt{s}}).\end{aligned} \tag{16}$$

Note that $\boldsymbol{s}^t_0$ represents a zero vector $\mathbf{0}$, and w^t_0 denotes an initial symbol. In an output layer of the decoder network, predicted probabilities of w^t_n are produced by:

$$\boldsymbol{o}^t_n = \texttt{SOFTMAX}(\boldsymbol{s}^t_n, \boldsymbol{\theta}_{\texttt{o}}), \tag{17}$$

where $\boldsymbol{o}^t_n$ means $P(w^t_n|\boldsymbol{s}^t_n, \boldsymbol{\Theta})$ and $\boldsymbol{\Theta}$ corresponds to $\{\boldsymbol{\theta}_{\texttt{w}}, \boldsymbol{\theta}_r, \boldsymbol{\theta}_{\texttt{u}}, \boldsymbol{\theta}_{\texttt{h}}, \boldsymbol{\theta}_{\texttt{s}}, \boldsymbol{\theta}_{\texttt{o}}\}$. While the RA-LSTM-RNN-LMs are trained from utterances with role labels, the RPDA-LMs are trained from multi-turn dialogues. The parameter can be optimized by:

$$\hat{\boldsymbol{\Theta}} = \operatorname*{argmin}_{\boldsymbol{\Theta}} - \sum_{(\boldsymbol{W}, \boldsymbol{R}) \in \mathcal{D}} \log P(\boldsymbol{W}|\boldsymbol{R}, \boldsymbol{\Theta}), \tag{18}$$

where $\mathcal{D}$ denotes the training data set that includes multiple multi-turn dialogues with role labels. The optimization is also followed by backpropagation through time.

4. Experiments

4.1. Setups

We used the contact center dialogue data sets, which include several topics. One dialogue means one telephone call between one operator and one customer. Each dialogue was separately recorded and the data set consists of 2,636 dialogues. One dialogue included about 121 utterances, and one utterance included about 10 words in average. We divided the data set into a training set, a validation set, and a test set. Vocabulary size of the training set was 25K words. Table 1 shows details.

Table 2: *PPL and WER (%) results.*

	Models		Role labels	Past utterances	Valid PPL	Valid WER	Test PPL	Test WER
(1).	HPY-LM	-	-	-	26.97	23.58	25.54	23.37
(2).	LSTM-RNN-LM	-	-	-	27.24	23.53	26.08	23.12
(3).	LSTM-RNN-LM	+ HPY-LM	-	-	22.55	22.03	21.60	21.87
(4).	RA-LSTM-RNN-LM	-	$\surd$	-	26.49	23.62	25.51	23.24
(5).	RA-LSTM-RNN-LM	+ HPY-LM	$\surd$	-	22.05	22.00	21.19	21.85
(6).	DC-LM ($\boldsymbol{h}_{t-1} = \mathbf{0}$)	-	-	-	25.67	23.04	24.66	22.71
(7).	DC-LM ($\boldsymbol{h}_{t-1} = \mathbf{0}$)	+ HPY-LM	-	-	21.43	21.85	20.58	21.79
(8).	DC-LM	-	-	$\surd$	21.32	22.15	20.50	21.68
(9).	DC-LM	+ HPY-LM	-	$\surd$	18.92	21.17	18.14	21.09
(10).	RPDA-LM ($\boldsymbol{h}_{t-1} = \mathbf{0}$)	-	$\surd$	-	25.22	22.85	24.11	22.62
(11).	RPDA-LM ($\boldsymbol{h}_{t-1} = \mathbf{0}$)	+ HPY-LM	$\surd$	-	21.34	21.79	20.41	21.74
(12).	RPDA-LM	-	$\surd$	$\surd$	19.40	21.92	18.63	21.44
(13).	RPDA-LM	+ HPY-LM	$\surd$	$\surd$	**17.56**	**21.10**	**16.83**	**20.95**

Table 1: *Experimental data set.*

	# of dialogues	# of words
Train	2,545	3,318,235
Valid	45	48,511
Test	46	50,116

We used a senone based LSTM-RNN acoustic model. Acoustic feature consisted of 40 dimensional log mel-filterbank coefficients appended with delta and acceleration coefficients, which frame shift was 10 ms. For the acoustic modeling, we stacked two LSTM-RNN layers with 512 cells, two fully connected layers which had 1,024 hidden units with rectified linear units, and a softmax layer with 3,072 outputs. The speech recognizer is a weighted finite state transducer (WFST) based decoder [32].

For evaluation, we constructed various LMs. "HPY-LM" is 3-gram hierarchical Pitman-Yor LM which is the baseline n-gram LM [2]. The HPY-LM was introduced in the speech recognizer by converting WFST. In addition, we constructed the following four neural LMs. "LSTM-RNN-LM" is an LSTM-RNN based LM and "RA-LSTM-RNN-LM" is the LM described in Section 2. In RA-LSTM-RNN-LM, a word continuous representation and a role continuous representation were set to 650 and 32 dimensional vectors, and LSTM-RNN layer had 650 units. LSTM-RNN-LM is composed by deleting role labels in RA-LSTM-RNN-LM. These two LMs only used contexts within an utterance. "DC-LM" is a document context LM that is modeled by hierarchical RNN encoder-decoder modeling, and "RPDA-LM" is the proposed method detailed in Section 3. In RPDA-LM, a word continuous representation and a role continuous representation were set to 650 and 32 dimensional vectors, LSTM-RNN layer in a word-level encoder network had 650 units, LSTM-RNN layer in a utterance-level encoder network had 200 units, and LSTM-RNN layer in a decoder network had 650 units. Note that the DC-LM is composed by deleting role labels in the RPDA-LM. These two LMs can capture relationships between utterances. For optimization of each neural LM, we used mini-batch stochastic gradient descent where a learning rate was scheduled using validation loss. The dropout rate in each LSTM-RNN layer was set to 0.5. For implementing neural LMs to ASR, 100-best rescoring was conducted. When using DC-LM and RPDA-LM, we incrementally rescored each utterance in order from the front because the LMs depend on multi-turn dialogues. In addition, we examined linear interpolation of HPY-LM and each neural LMs in which the interpolation weights were optimized using the validation set.

4.2. Results

Experimental results in terms of perplexity (PPL) and word error rate (WER) are shown in Table 2. In Table 2, "Role labels" and "Past utterances" represents that either role labels were utilized or not, and either past utterances were utilized or not in each LM. Line (1) shows baseline results in which no neural LMs were introduced. It is shown that PPL results in the contact center dialogue tasks were comparatively smaller than those in general ASR tasks because many backchannels are involved in both speaker's utterances. Results obtained with LSTM-RNN-LMs are shown in lines (2) and (3), and those obtained with RA-LSTM-RNN-LMs are shown in lines (4) and (5). They show that (3) outperformed (1) and (2), and (5) outperformed (3) and (4). This indicates that neural LMs can be complemented with the n-gram LM. In addition, (4) was comparable to (2). This indicates that role information was not so effective for estimating the generative probability of words within an utterance. Results obtained with the DC-LMs are shown in lines (6) to (9), and those obtained with the RPDA-LMs are shown in lines (10) to (13). Note that (6), (7), (10) and (11) show results where $\boldsymbol{h}_{t-1}$ in utterance-level encoder network was set to a zero vector $\mathbf{0}$. They show that (12) outperformed (10) and (8) outperformed (6). This indicates that it is important to consider past long-range utterances for improving word estimation performance. They also show that (12) was superior to (8). This indicates that role information is effective for capturing sequential interactive contexts. The best performance was obtained by (13), which achieved 2.42 point WER improvement compared to (1) and 0.92 point WER improvement compared to (3) in the test set.

5. Conclusions

In this paper, we described RPDA-LMs that can be used in role play dialogue ASR tasks. The main advantage of the RPDA-LM is that it leverages sequential interactive contexts in role play multi-turn dialogues for estimating the generative probability of words. We formulated the RPDA-LM as conditional hierarchical recurrent encoder-decoder modeling with multiple LSTM-RNNs. Experiments in a contact center dialogue task showed that the RPDA-LM could yield ASR performance improvements compared with RA-LSTM-RNN-LM and DC-LM. In addition, we verified the effectiveness of explicitly capturing both long-range sequential utterances and role labels.

6. References

[1] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech and Language*, vol. 13, pp. 359–393, 1999.

[2] Y. W. Teh, "A hierarchical Bayesian language model based on Pitman-Yor processes," *In Proc. Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING/ACL)*, pp. 985–992, 2006.

[3] R. Masumura, T. Asami, T. Oba, H. Masataki, S. Sakauchi, and A. Ito, "Combination of various language model technologies including data expansion and adaptation in spontaneous speech recognition," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 463–467, 2015.

[4] R. Masumura, T. Asami, T. Oba, H. Masataki, S. Sakauchi, and A. Ito, "Investigation of combining various major language model technologies including data expansion and adaptation," *IEICE Transaction on Information and Systems*, vol. E99-D, no. 10, pp. 2452–2461, 2016.

[5] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *Journal of Machine Learning Research*, vol. 3, pp. 1137–1155, 2003.

[6] H. Schwenk, "Continuous space language models," *Computer speech and Language*, vol. 21, pp. 492–518, 2007.

[7] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, and S. Khudanpur, "Recurrent neural network based language model," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1045–1048, 2010.

[8] T. Mikolov, S. K. Stefan, L. Burget, J. Cernocky, and S. Khudanpur, "Extensions of recurrent neural network language model," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5528–5531, 2011.

[9] M. Sundermeyer, H. Ney, and R. Schluter, "From feedforward to recurrent LSTM neural networks for language models," *IEEE/ACM Transactions of Audio, Speech and Language processing*, vol. 23, no. 3, pp. 517–529, 2015.

[10] R. Lin, S. Liu, M. Yang, M. Li, M. Zhou, and S. Li, "Hierarchical recurrent neural network for document modeling," *In Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 899–907, 2015.

[11] T. Wang and K. Cho, "Larger-context language modelling with recurrent neural network," *In Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1319–1329, 2016.

[12] G. Ji and J. Bilmes, "Multi-speaker language modeling," *In Proc. Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 133–136, 2004.

[13] H. Ashikawa, N. Tawara, A. Ogawa, T. Iwata, T. Kobayashi, and T. Ogawa, "Exploiting end of sentences and speaker alternations in language modeling for multiparty conversations," *In Proc. Asia-Pacific Signal and Information Processing Association (APSIPA)*, 2017.

[14] Y.-C. Tam and P. Vozila, "Unsupervised latent speaker language modeling," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1477–1480, 2011.

[15] R. Masumura, T. Oba, H. Masataki, O. Yoshioka, and S. Takahashi, "Role play dialogue topic model for language model adaptation in multi-party conversation speech recognition," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4873–4877, 2014.

[16] B. Liu and I. Lane, "Dialogue context language modeling with recurrent neural networks," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5715–5719, 2017.

[17] A. Sordoni, Y. Bengio, H. Vahabi, C. Lioma, J. G. Simonsen, and J.-Y. Nie, "A hierarchical recurrent encoder-decoder for generative context-aware query suggestion," *In Proc. ACM International on Conference on Information and Knowledge Management (CIKM)*, pp. 553–562, 2015.

[18] O. Vinyals and Q. Le, "A neural conversational model," *In Proc. International Conference on Machine Learning (ICML) Deep Learning Workshop*, 2015.

[19] L. Shang, Z. Lu, and H. Li, "Neural responding machine for short-text conversation," *In proc. Annual Meeting of the Association for Computational Linguistics and International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pp. 1577–1586, 2015.

[20] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A diversity-promoting objective function for neural conversation models," *In Proc. Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 110–119, 2016.

[21] J. Li, M. Galley, C. Brockett, G. P. Spithourakis, J. Gao, and B. Dolan, "A persona-based neural conversation model," *In Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 994–1003, 2016.

[22] H. Mei, M. Bansal, and M. R. Walter, "Coherent dialogue with attention-based language models," *In Proc. AAAI Conference on Artificial Intelligence (AAAI)*, pp. 3252–3258, 2017.

[23] R. Lowe, N. Pow, I. V. Serban, L. Charlin, C.-W. Liu, and J. Pineau, "Training end-to-end dialogue systems with the ubuntu dialogue corpus," *Dialogue & Discourse*, vol. 1, pp. 31–65, 2017.

[24] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," *In Proc. AAAI Conference on Artificial Intelligence (AAAI)*, pp. 3776–3783, 2016.

[25] I. V. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. Courville, and Y. Bengio, "A hierarchical latent variable encoder-decoder model for generating dialogues," *In Proc. AAAI Conference on Artificial Intelligence (AAAI)*, pp. 3295–3301, 2017.

[26] T. Mikolov and G. Zweig, "Context dependent recurrent neural network language model," *In Proc. Spoken Language Technology Workshop (SLT)*, pp. 234–239, 2012.

[27] X. Chen, T. Tan, X. Liu, P. Lanchantin, M. Wan, M. J. F. Gales, and P. C. Woodland, "Recurrent neural network language model adaptation for multi-genre broadcast speech recognition," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 3511–3515, 2015.

[28] S. Deena, R. W. M. Ng, P. Madhyastha, L. Specia, and T. Hain, "Semi-supervised adaptation of RNNLMs by fine-tuning with domain-specific auxiliary features," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 2715–2719, 2017.

[29] S. Toyama, D. Saito, and N. Minematsu, "Use of global and acoustic features associated with contextual factors to adapt language models for spontaneous speech recognition," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 543–547, 2017.

[30] C. Hori, T. Hori, S. Watanabe, and J. R. Hershey, "Context-sensitive and role-dependent spoken language understanding using bidirectional and attention LSTMs," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 3236–3240, 2017.

[31] N. Sawada, R. Masumura, and H. Nishizaki, "Parallel hierarchical attention networks with shared memory reader for multi-stream conversational document classification," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 3311–3315, 2017.

[32] T. Hori, C. Hori, Y. Minami, and A. Nakamura, "Efficient WFST-based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition," *IEEE transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1352–1365, 2007.